Pacific Northwest
NATIONAL LABORATORY

Building Protein-Protein Interaction Networks from Relational Databases

August 24, 2023

SJ Kim

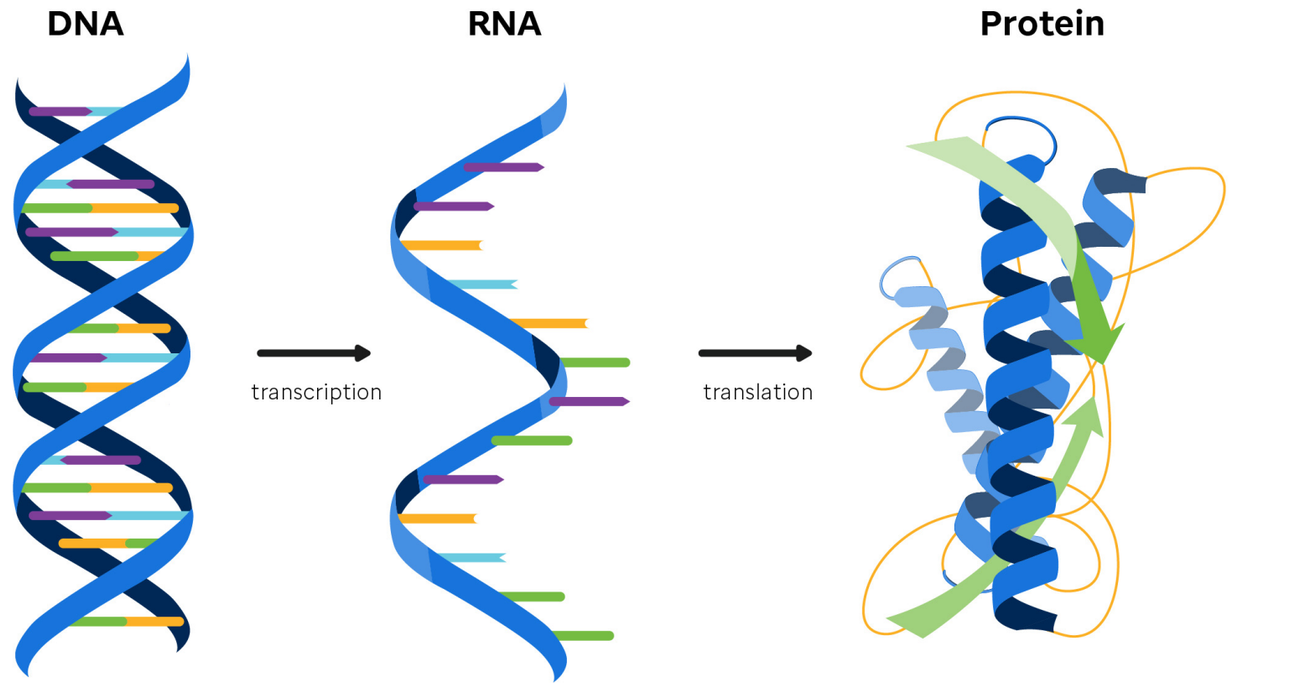
NSIP Master's Intern
University of Oregon

Mentors Dr. Lisa Bramer and David Degnan

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Proteins are the Molecular Workhorses of Cellular Processes



genomics

proteomics

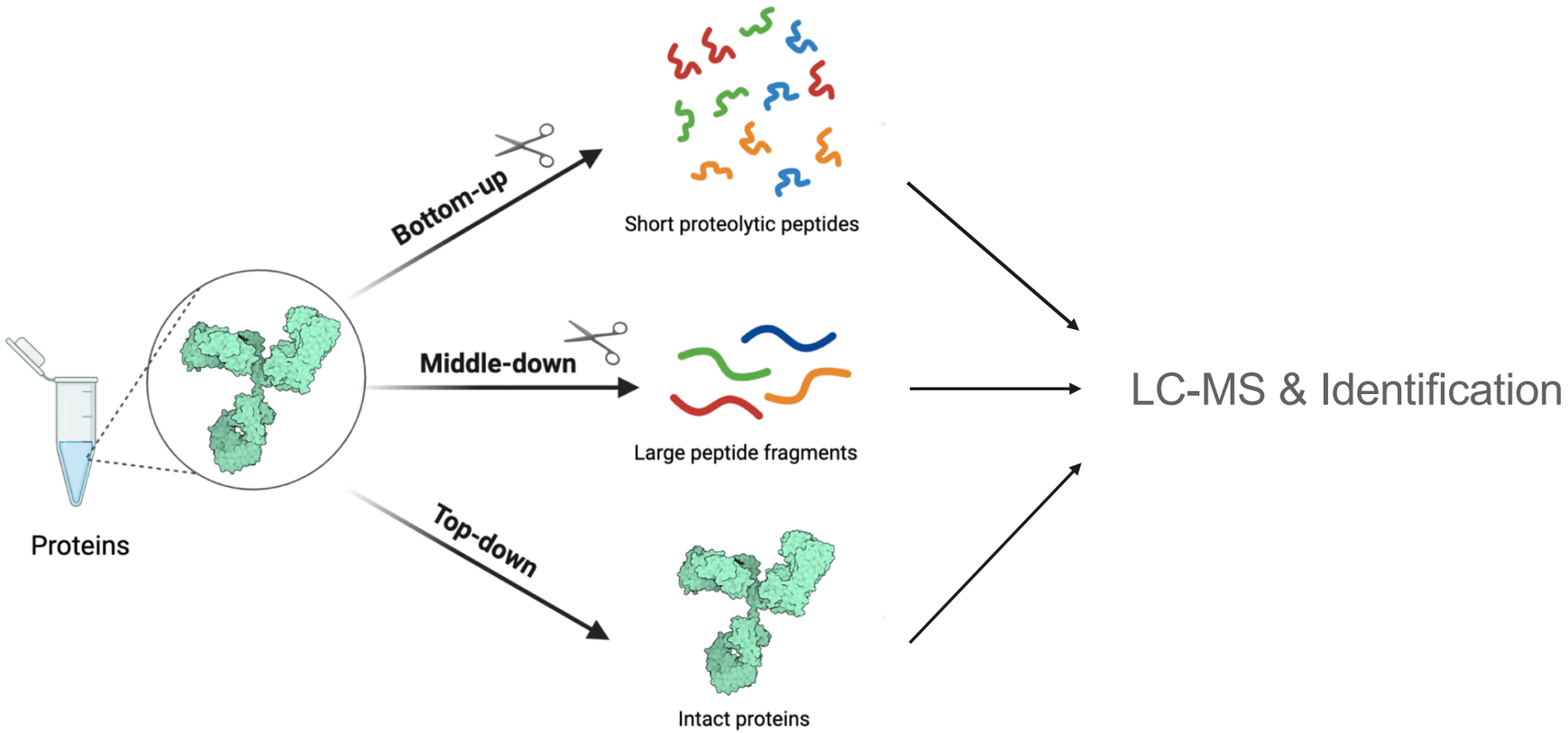
- DNA Replication
- Communication between cells
- DNA Repair
- Cell life cycle
- Gene Expression
- Forming structural elements

- Protein-Protein Interaction (PPI) is the foundation for understanding cells, cellular processes, and how they differ in different biological systems and states (like healthy vs. disease)

https://learngenomics.dev/docs/biological-foundations/the-central-dogma/

Improvements in Mass Spectrometry has led to Exponential Growth of Proteomics Data

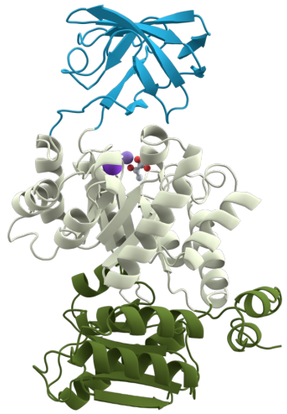
- Mass Spec: used to identify proteins

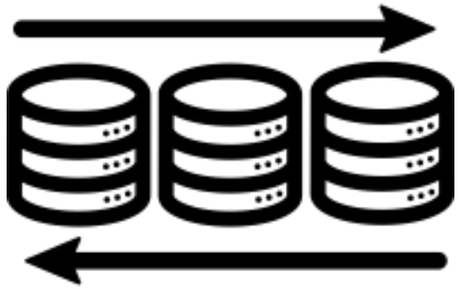


https://www.rapidnovor.com/what-is-protein-mass-spectrometry/

Project Question, Solution, and End Goal

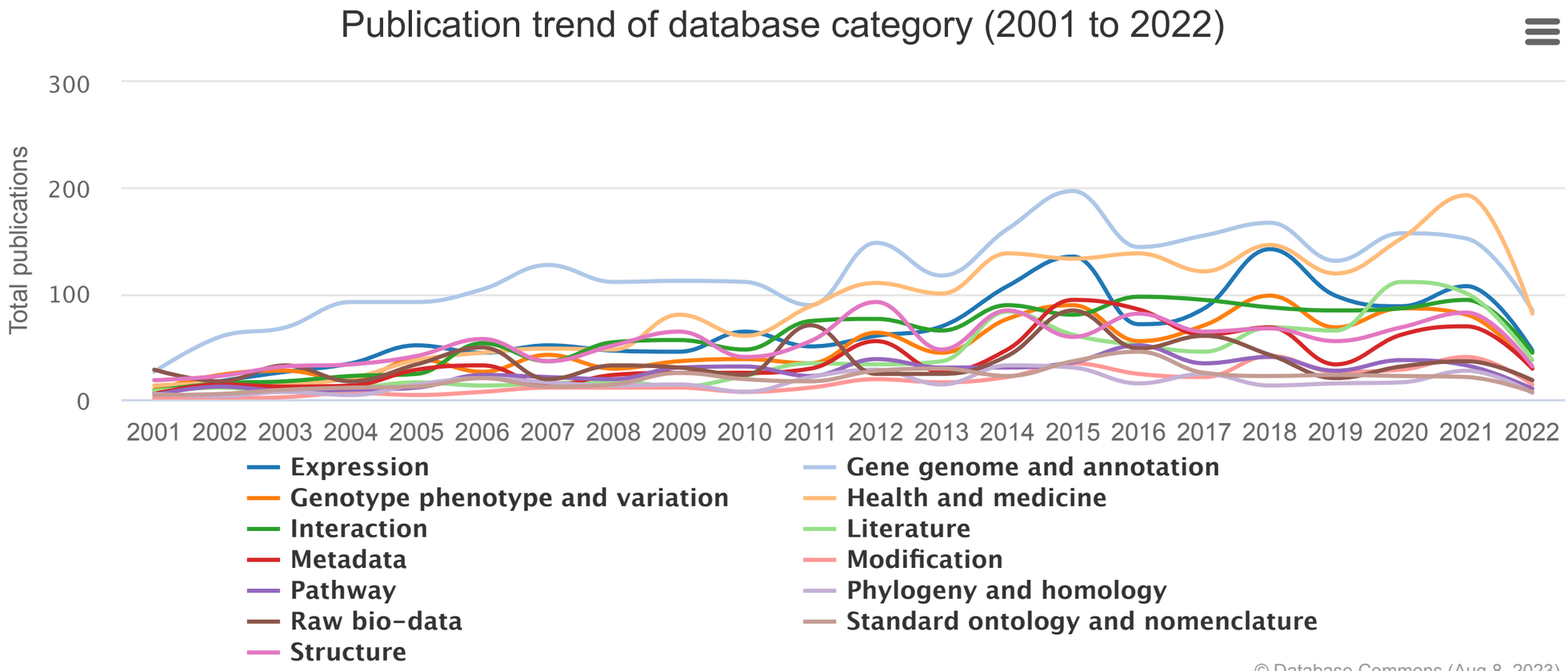
- **Question:** How can we use existing databases of hand-annotated proteins and protein interactions?
- **Solution:** Build a pipeline that researchers can use to query organisms of interest to quickly obtain network of verified PPI.
- **End Goal:** These networks will be used as a basis for the prior distributions in a Bayesian graphical model. The evidence for the Bayesian model (used to generate posterior distributions) will be the relational data from a specific experiment.





Pacific Northwest
NATIONAL LABORATORY

Literature Review of PPI Databases



© Database Commons (Aug 8, 2023)
National Genomics Data Center

5,900+ biological databases worldwide

Literature Review of PPI Databases

- Criteria of interest: experimentally verified protein-protein interactions, regularly updated/maintained, broad array of species, programmatic access

	Knowledge Base	Protein-Protein Interaction Network Available	Experimentally Verified Interactions	Inter-Species Relationship Available		
2	DAVID					Legend
3	KEGG					Yes
4	cBioPortal					No
5	STRING			*Pulls from IntAct		Unknown
6	ENCODE					
7	UniProt			*Pulls from IntAct		
8	IGSR					
9	InterPro					
10	SILVA					
11	gnomAD					
12	GO					
13	NCBI					
14	BioGRID					
15	Reactome					
16	Metacyc					
17	IntAct					

Pacific Northwest NATIONAL LABORATORY

IntAct

IntAct

Home | Download | About | Documentation | Feedback

IntAct Molecular Interaction Database

IntAct provides a free, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions. The IntAct Team also produces the Complex Portal. You are currently visiting the new website of IntAct. The former version can be found here and will be supported until the end of 2023.

IntAct's COVID-19 dataset

miXML$_{2.5}$ | miXML$_{3.0}$

The data primarily covers protein-protein and several RNA-protein interactions involving SARS-CoV2 and SARS-CoV. All interactions from the relevant publications are covered in this dataset, including interactions with other organism.

Quick Search | Batch Search | Advanced Search

Search by gene names, UniProt ACs, Pubmed, protein names, Complex ACs

Examples:

- Gene names: Ndc80
- UniProt ACs: Q05471
- Taxon IDs: 9606
- Publication IDs: 32353859
- Complex ACs: CPX-5742
- GO terms: GO:0016491

Quick search can yield to false positives as one identifier can be referenced by 2 different fields. If you want to avoid these, we suggest you to use Advanced Search.

Newsletter

email address | Subscribe

Featured Dataset

Large-scale phosphomimetic screening identifies phospho-modulated motif-based protein interactions - **Kliche J et al.**

Access | Download

Archive

Latest News

IntAct Portal version: 1.0.3 - December 2021

Release 244 - July 2023

- Publications : 23,170
- Interactors : 134,292
- Interactions : 804,367
- Binary Interactions : 1,263,084

Tweets

Pipeline Structure

Approaches that didn't work:

- Pulling from STRING/UniProt to access IntAct data
- Using a ChromeDriver to scrape webpages for data

Pacific Northwest NATIONAL LABORATORY

Example Organism: Yeast

	PMID	ParticipantA	ParticipantB	SpeciesA	TaxonID_A	SpeciesB	TaxonID_B
2	10627553	P28006	P20604	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
3	10627553	P28006	P20604	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
4	11208108	P52917	P25604	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
5	11208108	P36095	P25604	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
6	11470436	P10507	P11914	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
7	11470436	P11914	P10507	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
8	11283612	P47104	P52286	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
9	11283612	P32324	P52286	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
10	11283612	P38352	P52286	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
11	11283612	P38352	Q12018	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
12	11283612	P17255	P47104	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
13	11283612	P47104	P16140	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
14	11283612	P22203	P47104	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
15	11283612	P47104	P32610	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
16	11283612	P16140	Q03956	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
17	11283612	Q03956	P32610	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
18	11733989	P53829	P25655	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
19	11733989	P25655	P53280	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
20	11733989	P53829	P22204	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
21	11733989	P53829	P34909	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
22	11733989	P53280	P25655	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
23	11733989	P53280	Q12514	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
24	11733989	P53280	P25655	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
25	11036083	Q12432	Q08649	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
26	11036083	Q12432	Q08649	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
27	11036083	Q12432	Q08649	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
28	11726501	P26309	P26449	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
29	11726501	P26309	P26449	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
30	10913169	P20438	P00546	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
31	10913169	P13365	P00546	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
32	10913169	P20438	P00546	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292
33	11545742	P15873	Q04049	Saccharomyces cerevisiae	559292	Saccharomyces cerevisiae	559292

Conclusion and Next Steps

- Lots of high-quality experimentally curated information about PPI out there
- No tool/source/method that allows a user to query by organism for PPI
- Widely variable PPI data: metadata, identifiers, organization, etc.
- Next Steps
  - Automate FTP download
  - Wrap script to pull queried organisms “on the fly”
  - Fold script into network – to be used as prior distributions for a Bayesian model

Pacific Northwest
NATIONAL LABORATORY

Acknowledgements

David Degnan
Dr. Lisa Bramer
Moses Obiri
Funding: Predictive Phenomics Initiative
Office of STEM Edu

PPI
PREDICTIVE PHENOMICS
INITIATIVE
@PNNL

Pacific Northwest
NATIONAL LABORATORY

Thank you

Pacific Northwest
NATIONAL LABORATORY

Questions?

SJ Kim
NSIP Master's Intern

sj.kim@pnnl.gov